Alineamiento Múltiple de Secuencias

Curso de Algoritmos en Bioinformática

Luis Garreta

Doctorado en Ingeniería
Pontificia Universidad Javeriana – Cali

April 7, 2018

Alineamientos de Secuencias

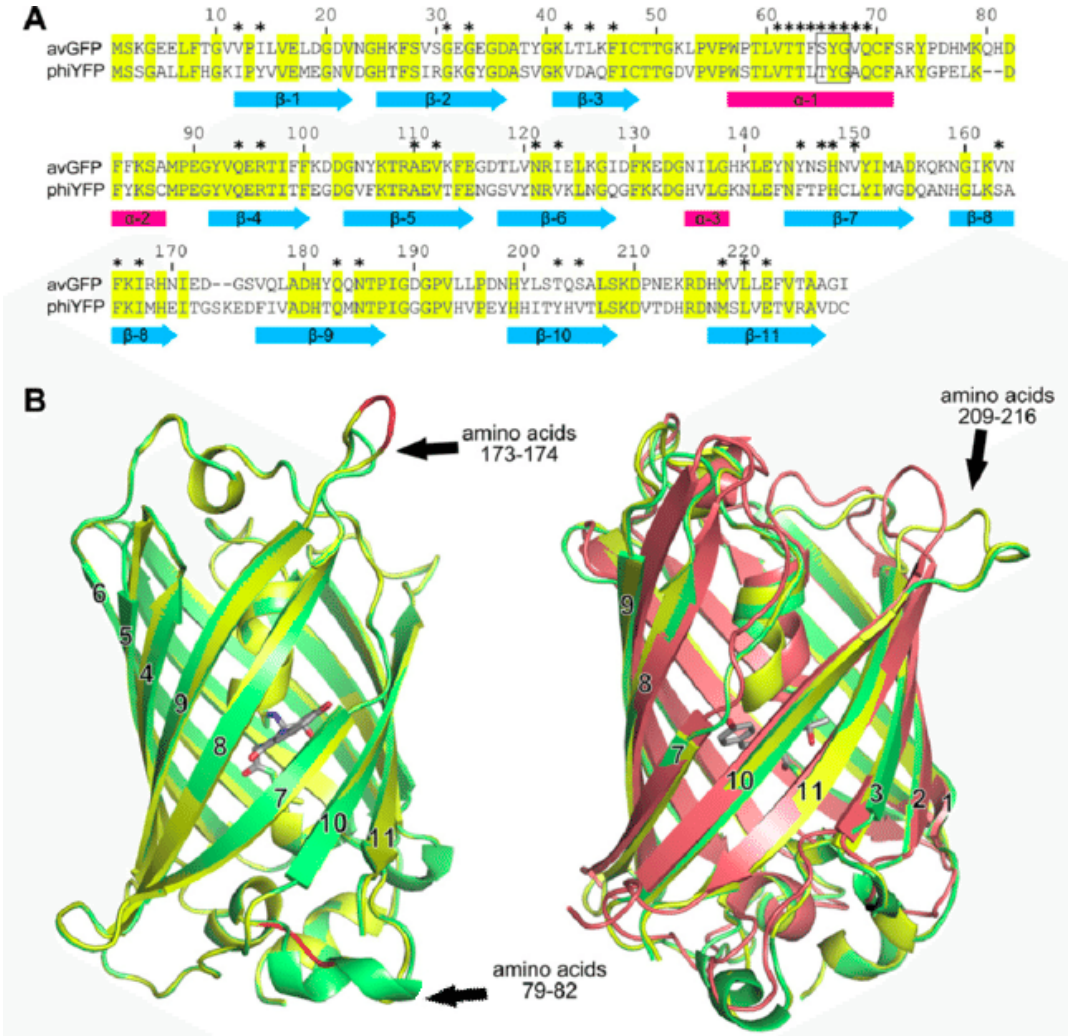
- En bioinformática un alineamiento de secuencias es una forma de organizar las secuencias (ADN, ARN, Proteínas) para identificar **regiones similares** que posiblemente tienen una **relación biológica**.

Relaciones entre secuencias

- Las regiones similares entre secuencias pueden mostrar que existe relaciones biológicas como:
  - ✓ **Relaciones funcionales**: realizan la misma función
  - ✓ **Relaciones estructurales**: tienen la misma estructura
  - ✓ **Relaciones evolutivas:** tiene un ancestro común

Ejemplo: Alineamiento de secuencias de dos proteínas

Regiones similares corresponden a estructuras comunes en las dos proteínas



Alineamientos Múltiples de Secuencia (AMS o MAS)

Un AMS es un alineamiento de más de 2 secuencias (n $>$ 2).

Alineamientos Múltiples de Secuencia (AMS o MAS)

Un AMS es un alineamiento de más de 2 secuencias (n $>$ 2).

- Un grupo de secuencias puede tener algunas regiones altamente conservadas:

Alineamientos Múltiples de Secuencia (AMS o MAS)

Un AMS es un alineamiento de más de 2 secuencias (n $>$ 2).

- Un grupo de secuencias puede tener algunas regiones altamente conservadas:

- Los algoritmos de AMS buscan detectar estas regiones y alinearlas asumiendo que son comunes en las secuencias.

Características de las secuencias en un AMS

- Se asume que:
  - ✓ Las secuencias tienen la misma longitud
  - ✓ Las secuencias tienen alguna relación biológica:
    - funcional,
    - estructural,
    - evolutiva,
    - otra.
  - ✓ Las secuencias han sufrido mutaciones.

Resultado de un Alineamiento Múltiple de Secuencias

- El resultado es una matriz de LxN donde L es la longitud de la secuencia y N es el número de secuencias
- Los residuos o regiones comunes están alineados en lo posible en posiciones similares o cercanas

Propósito de los AMS?

- Los AMS sirven de base para realizar muchos otros análisis bioinformáticos, entre las principales

Extrapolación	Para determinar la familía de proteínas de una secuencia desconocida
Análisis filogénetico	Reconstruir la historia evolutiva de un conjunto de proteínas a través de un árbol filogenético
Identificación de Perfiles o Motivos	Detección de regiones conservadas pequeñas (motivos) o grandes (dominios) que se repiten en todas en una o más secuencias
Creación de perfiles	Es posible definir un **perfil** cuando elementos específicos se repiten en la misma posición de las secuencias.
Predicción de estructura	Un buen alineamiento puede mostrar estructuras secundarias que se repiten

Tipos de Algoritmos para AMS

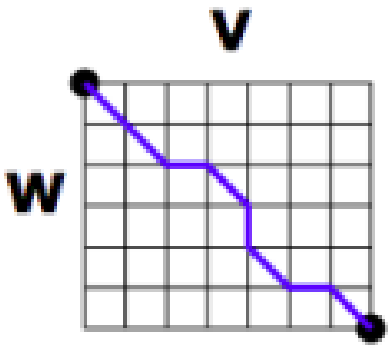
- **Algoritmos Exactos (Programación Dinámica)**
  Calculan un alineamiento óptimo para una función de puntaje dada. Muy constosos computacionalmente, poco usados.

- **Algoritmos por Alineamientos Progresivos**
  Se basa en alineamientos de pares de secuencias. Alínea dos secuencias, dos alineamientos o un alineamiento con una secuencia. Muy rápidos, pero pueden propagar errores iniciales.

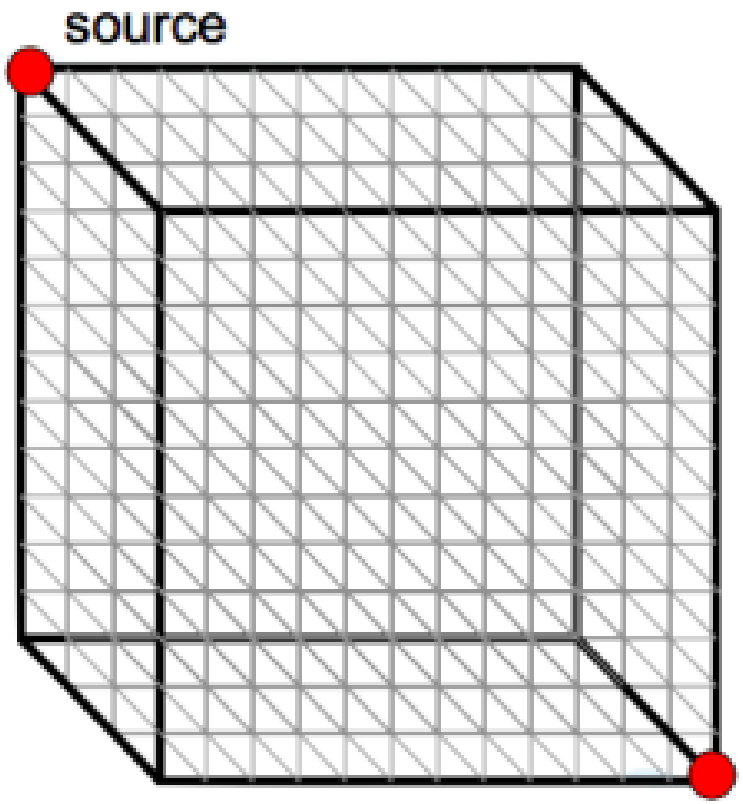
- **Algoritmos Alineamientos Iterativos**
  Trabajan de forma similar a los progresivos pero repetidamente realinean las secuencias iniciales así como van adicionando nuevas secuencias en el proceso. Muy buenos, un poco lentos.

Algoritmos Exactos con Programación Dinámica

- Similar al algoritmo para 2 secuencias sino que ahora son mas.
- Ejemplo N=3 secuencias con el algoritmo de Needleman-Wunsh

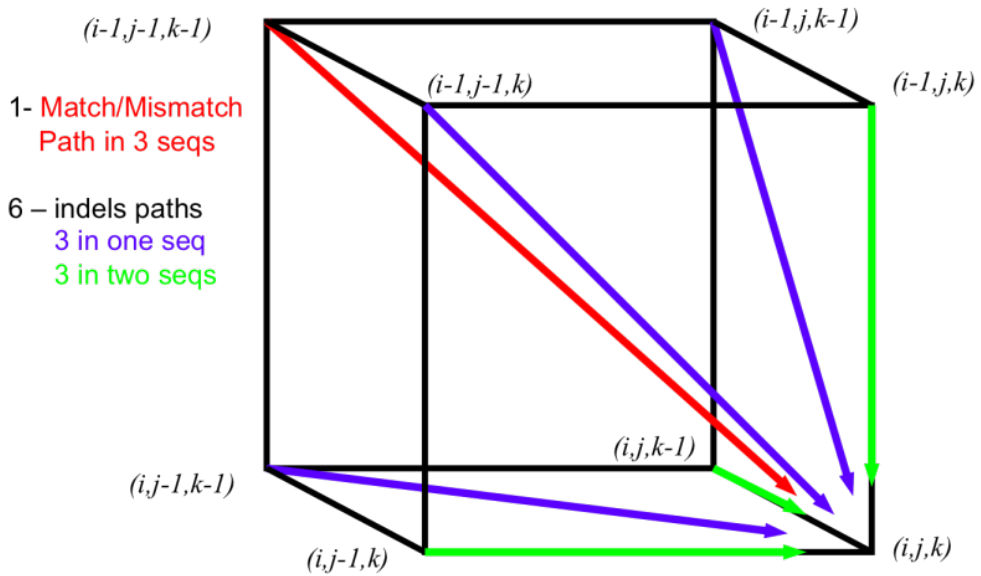


Matriz 2D



Matriz 3D

Algoritmos Exactos: Calculo de Puntajes



- Para 3 secuencias la complejidad es del orden $O(n^3)$
- Para k secuencias la complejidad es $O(2^k n^k)$
- Obtiene el alineamiento óptimo pero es inpractico computacionalmente

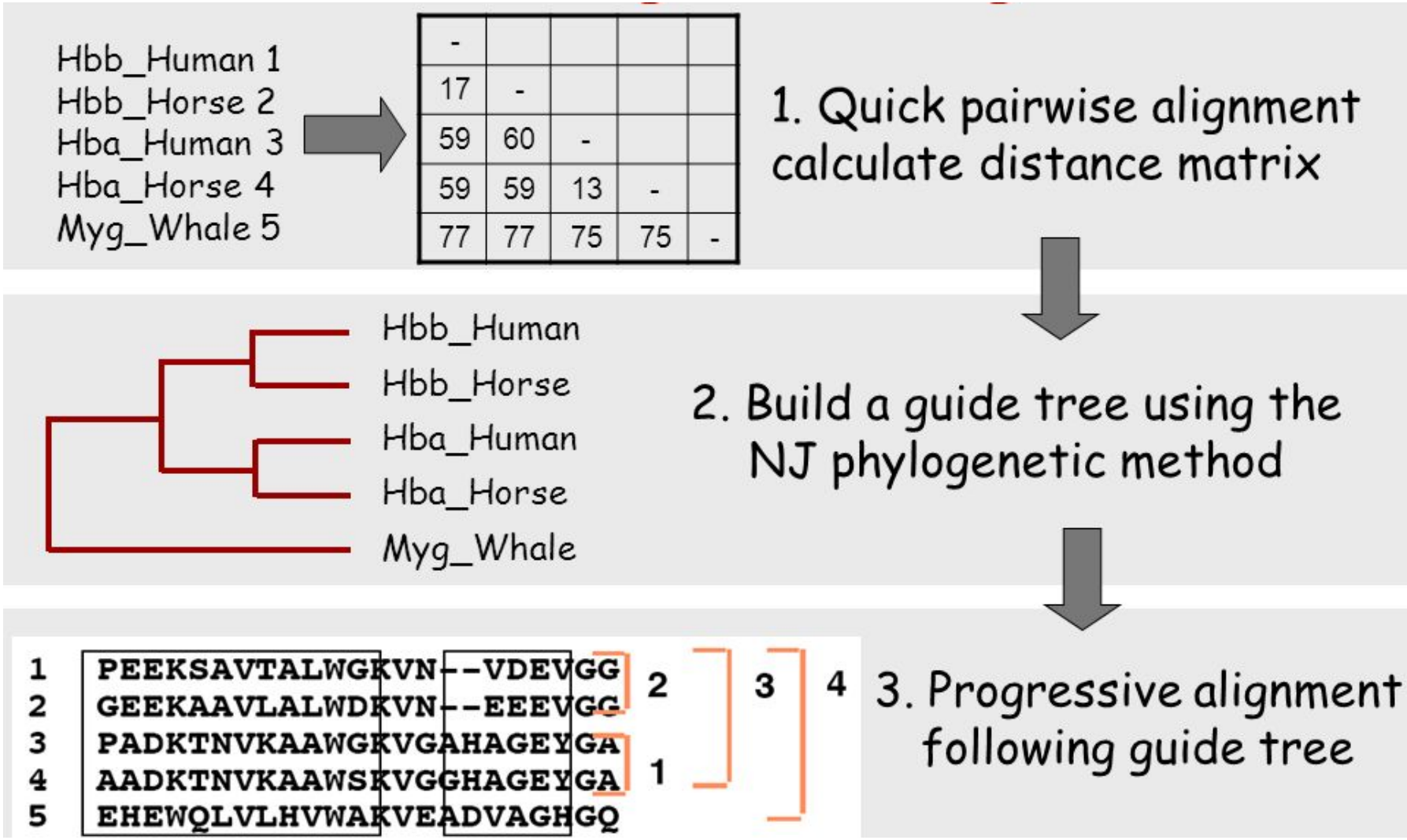
Métodos de Alineamiento Progresivos

Métodos de Alineamiento Progresivos

Este método va ensamblando progresivamente alineamientos de pares para formar un AMS

- Primero se lleva a cabo un alineamiento global de pares de secuencias:
  - ✓ Usando el algoritmo de Needleman-Wunsch
- Con los resultados se crea una matriz de distancias:
  - ✓ Está permite ver la relación evolutiva de la secuencia con las demás
- Se realiza un análisis filogenético simple,
  - ✓ Se crea un árbol filogenético (árbol guía)
  - ✓ Este árbol refleja la proximidad entre todas las secuencias
- El arbol guía es empleado para realizar un reajuste de las secuencias
  - ✓ Las dos secuencias más relacionadas son realineadas
  - ✓ Y se convirtién en un secuencia (consenso o **perfil**)
- Este proceso se continúa hasta que todas las secuencias quedan alineadas

Alineamiento Progresivo de CLUSTALW



Perfiles a Partir de un AMS

- **Qué es un Perfil:**
  - ✓ Es una tabla que lista las frecuencias de cada aminoácido en cada posición de la secuencia
  - ✓ Las frecuencias son calculadas a partir de un alineamiento múltiple
  - ✓ Representan (resumen) la estructura de un conjunto de secuencias
  - ✓ La tabla es la estructura (más estadística)
  - ✓ Se puede obtener una secuencia consenso (más visual)

Alineamiento Múltiple

1	2	3	4	5
K	L	M	–	K
K	L	K	L	K
K	M	M	L	–
M	L	–	L	M

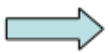
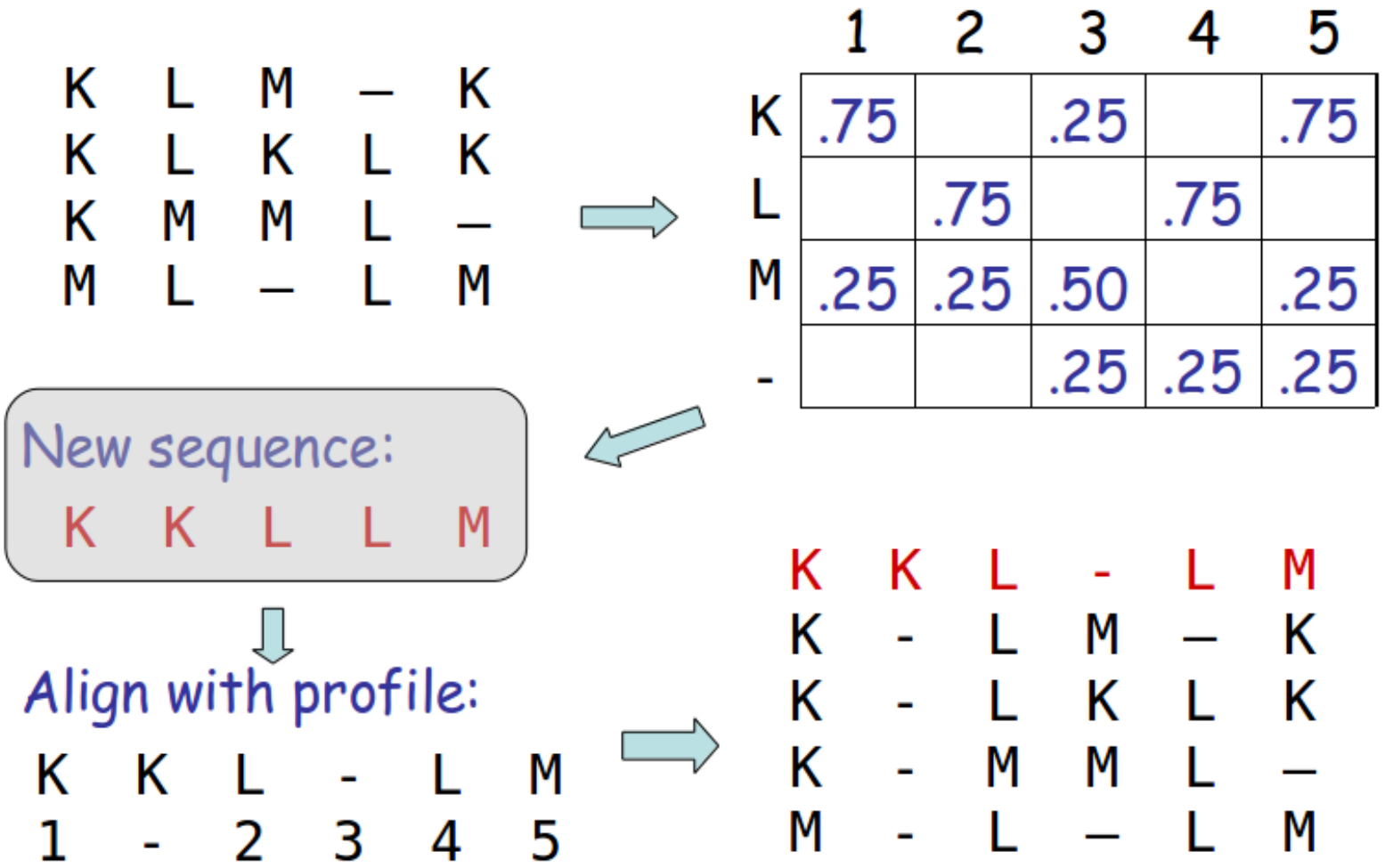


Tabla de Frecuencias

	1	2	3	4	5
K	.75		.25		.75
L		.75		.75	
M	.25	.25	.50		.25
-			.25	.25	.25

Secuencia Consenso **K L M L K**

Alineamiento de una secuencia a un perfil



Algoritmo de Alineamiento Progresivo

```
Algorithm 1: Método de alineamiento progresivo
Data: N secuencias
Result: Alineamiento de las N secuencias
1 begin
2    Construir la matriz de distancias
     /* Árbol guía */
3    Construir el árbol guía usando Neighbor-Joining
4    while no estén alineadas todas las secuencias do
5      Alinear las secuencias más relacionadas
6      Reducir las secuencias alineadas (Consensos, perfiles)
7    end
8 end
```

Ejemplo cinco globinas muy conocidas, bastante distantes

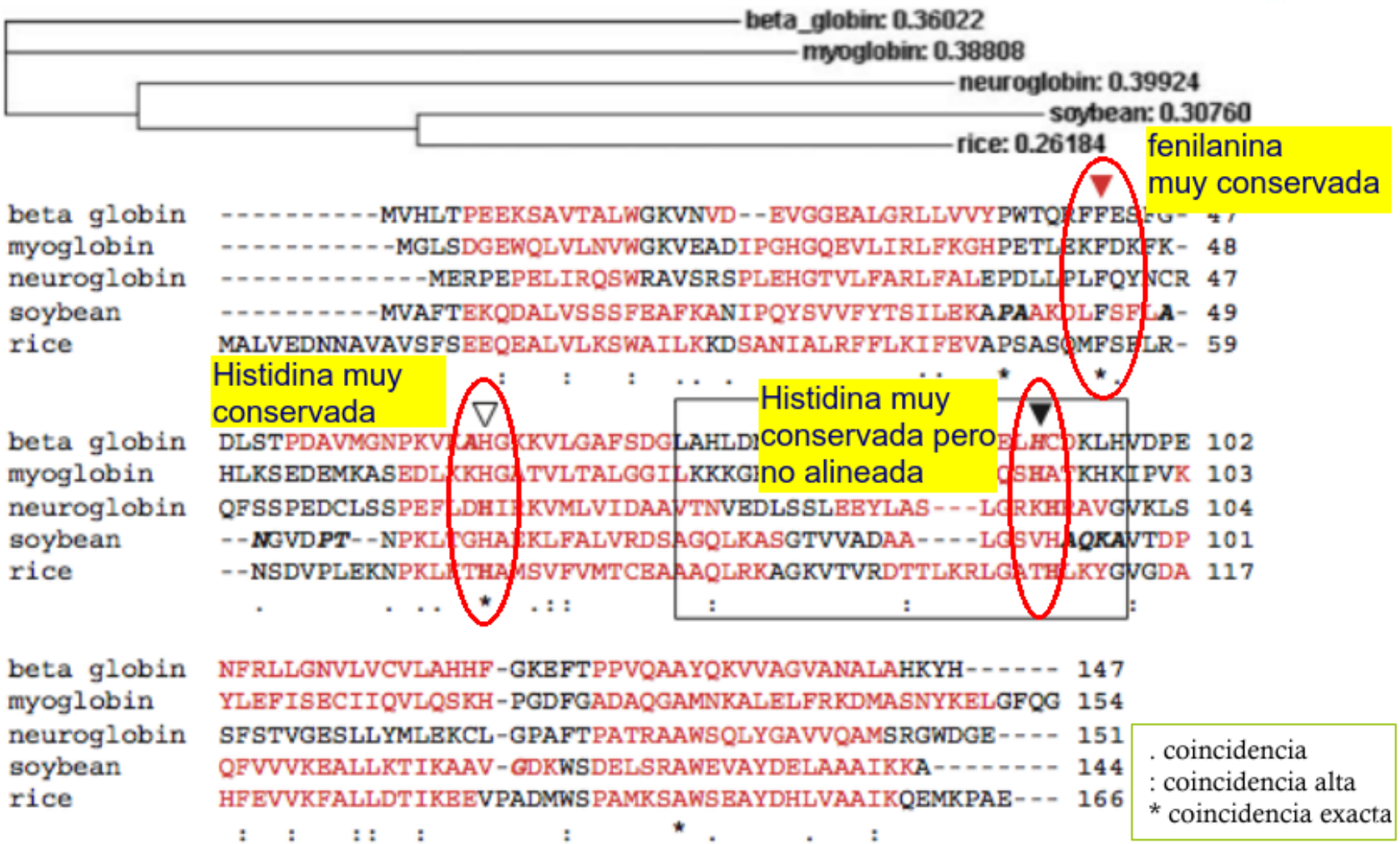
Creación del arbol gúia

- La longitud de las ramas depende de las distancias
- Se unen las ramas de las secuencias con distancias más cortas

```
(
beta_globin:0.36022,
myoglobin:0.38808,
(
neuroglobin:0.39924,
(
soybean:0.30760,
rice:0.26184)
:0.13652)
:0.06560);
```

Formato Newick (.nwk)

Ejemplo cinco globinas muy conocidas, bastante distantes



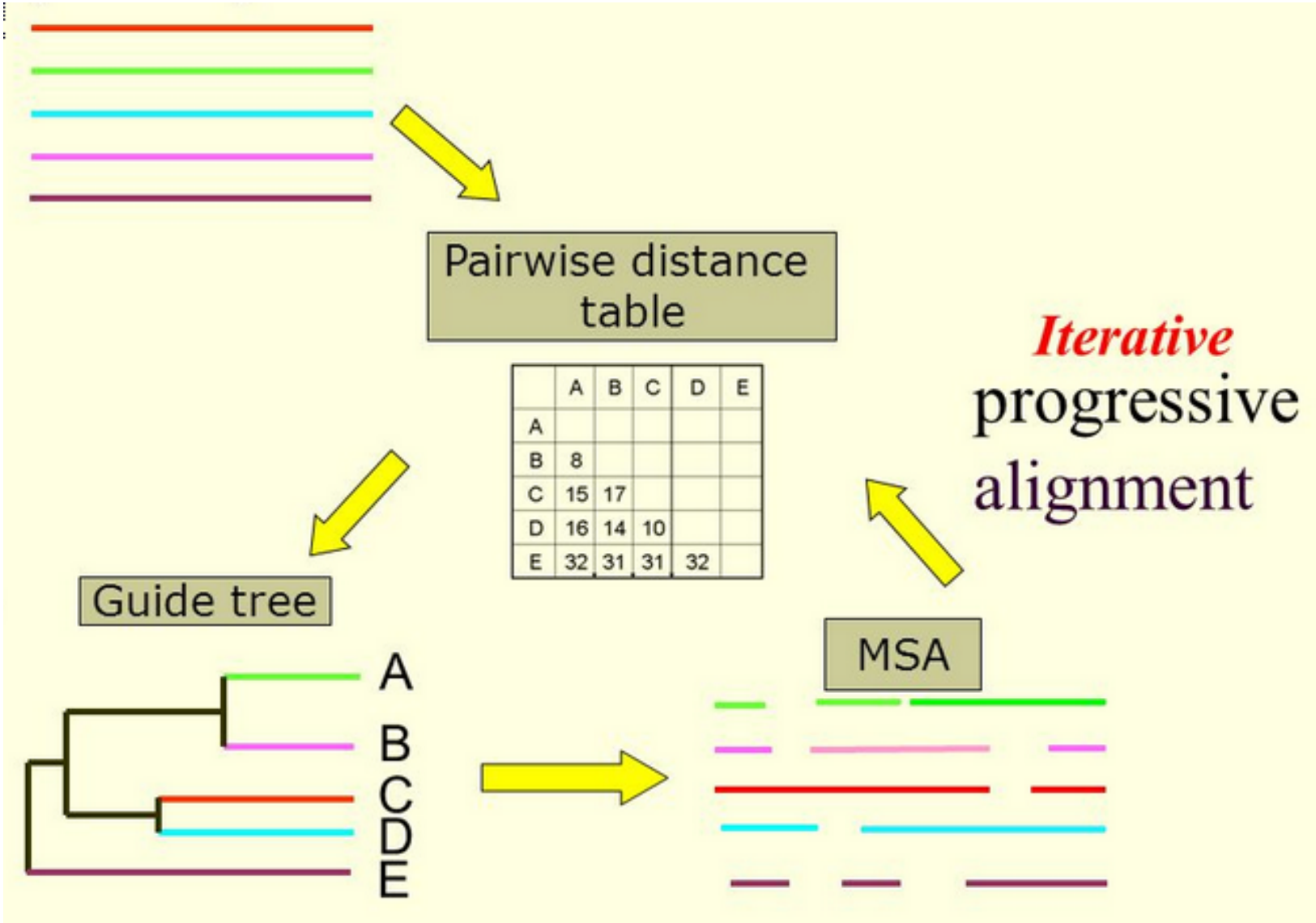
- Ejercicio: realizar este alineamiento múltiple mediante el ClustalW

Limitantes de los Alineamientos Progresivos

- Este método no es adecuado para comparar secuencias de diferentes longitudes (global)
- El resultado final proporcionado por éste método puede estar muy influenciado por el orden de las secuencias
- Debido a la naturaleza voraz (greedy algorithm) el resultado depende del alineamiento inicial de pares de secuencias (propagación de errores)
  - ✓ Si las dos primeras secuencias son muy similares, el alineamiento base contendrá pocos errores
  - ✓ Si las dos secuencias son muy divergentes los errores y los huecos se irán propagando

Metodos de Alineamiento Iterativos

Metodos de Alineamiento Iterativos



	A	B	C	D	E
A					
B	8				
C	15	17			
D	16	14	10		
E	32	31	31	32	

Algoritmo Iterativo de MUSCLE

- Actualmente, MUSCLE es una de los algoritmos más usados y que calcula buenos alineamientos
- MUSCLE, idea base: va revisando el árbol guía.

1. Parte como Clustal realizando alineamiento progresivo:

   1.1 Alineamiento de Pares
   1.2 Calculo de Distancias
   1.3 Construcción arbol guía
   1.4 Alineamiento y reajuste de las secuencias

2. Sobre la marcha va revisando si acaso el árbol guía ha cambiado y acaso alguna parte se puede realinear mejor.

   2.1 Calcula la similaridad de los alineamientos de a pares inducidos por el actual alineamiento múltiple.
   2.2 Si eso da un árbol distinto al que está usando, realínea según eso, y ve si el puntaje general mejora.

Tareas

1. (Jan) Estudiar el enfoque de Alineamientos Múltiples con algoritmos progresivos (Algoritmos Genéticos).
2. (Veronica) Realizar los alineamientos múltiples de las ejemplos mostrados utilizando:
   - 2.1 ClustalW
   - 2.2 Clustal-Omega
   - 2.3 MUSCLE
   - 2.4 T-Coffee
3. (???) Estudiar el algoritmo de Alineamiento Múltiple implementado por T-Coffee.

Referencias

- Xion (2006). Essential Bioinformatics (2006)
- Fahad Saeed and Ashfaq Khokhar. An Overview of Multiple Sequence Alignments .